Original Paper

Diagnosing Rare Genetic Diseases From Facial Images Using Visual Prompting With Multimodal Large Language Models: Diagnostic Accuracy Study

Jie Song, PhD; Zhichuan Xu, BS; Meng Xiao, BS; Cheng Bi, PhD; Yuxin Zhang, PhD; Xin Zheng, PhD; Xiaoran Li, BS; Ziyu Lu, BS; Qiongfang Cao, PhD; Bairong Shen, PhD

Department of Dermatovenereology and Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular, West China Hospital of Sichuan University, Chengdu, Sichuan, China

Corresponding Author:
Bairong Shen, PhD
Department of Dermatovenereology and Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular
West China Hospital of Sichuan University
No.2222, Xinchuan Road
Gaoxin District
Chengdu, Sichuan, 610212
China
Phone: 86 15995854635
Fax: 86 2861528682
Email: bairong.shen@scu.edu.cn

Abstract

Background: Rare genetic diseases often present with recognizable facial phenotypes. Whether multimodal large language models (MLLMs) can prioritize the diagnosis of rare genetic diseases from facial images and whether image enhancement or visual prompting can improve this process remain unclear.

Objective: This study aimed to evaluate the diagnostic accuracy of MLLMs for rare genetic diseases using facial images and to assess whether super-resolution enhancement or red-circle visual prompting improved performance.

Methods: This diagnostic accuracy study used 4746 facial images from the GestaltMatcher Database (GMDB), including 882 images in GMDB-Frequent from diseases with at least 6 images and 3864 images in GMDB-Rare from diseases with fewer than 6 images. Three locally deployed open-weight MLLMs—Gemma-4-31B, Qwen3.6-35B-A3B, and Qwen3.8-27B—generated 50 ranked candidate diagnoses under raw image, CodeFormer-enhanced, Qwen-Image-Edit-2511–generated red-circle–prompted, mismatched red-circle, and 2-stage image-to-phenotype conditions. Candidate diagnoses were mapped to Mondo Disease Ontology (MONDO) identifiers. In addition, we evaluated a random baseline, visual-prompt quality, expert-annotated regions, repeated-inference stability, model-reported confidence, diagnostic-list perplexity, prediction overlap, and disease composition.

Results: Top-50 accuracy ranged from 13.95% to 18.25% in GMDB-Frequent and from 0.57% to 1.35% in GMDB-Rare. Red-circle prompting produced numerical accuracy increases in 5 of 6 model-subset comparisons. Qwen3.8-27B showed a significant improvement in GMDB-Rare, with accuracy increasing from 0.57% to 1.27% (adjusted P=.009). In the good visual-prompt subgroup of GMDB-Frequent, Gemma-4-31B showed significant improvements over the super-resolution condition in both physician-rated analyses, with accuracy increases of 5.76 percentage points (Holm-adjusted P=.005) and 4.46 percentage points (Holm-adjusted P=.048), respectively. Matched red-circle inputs yielded higher accuracy than mismatched inputs across all model-subset comparisons, with adjusted P>.05. The image-to-phenotype workflow improved accuracy for Gemma-4-31B in GMDB-Frequent and Qwen3.8-27B in GMDB-Rare. In GMDB-Rare, the random baseline achieved 8.15% accuracy and exceeded all 3 models under the raw image and red-circle conditions. Prediction sets remained substantially overlapping after patient-image replacement, and top-1 predictions concentrated on a small number of diseases.

Conclusions: Under the evaluated experimental settings, the 3 MLLMs showed limited ability to diagnose rare genetic diseases from facial images, particularly diseases in GMDB-Rare. Visual prompting and image-to-phenotype diagnosis produced benefits in specific models and subgroups. The prediction-overlap, mismatched-image, and disease-composition analyses indicate that model outputs were only weakly constrained by patient-specific facial information. Further studies should evaluate physician comparisons, prospective multicenter data, and phenotype-grounded multimodal diagnostic workflows.

XSL•FO
RenderX

(JMIR Med Inform 2026;14:e105860) doi: 10.2196/105860

KEYWORDS

rare genetic diseases; facial phenotypes; multimodal large language models; visual prompting; artificial intelligence; AI

Introduction

The combined number of people affected by rare genetic diseases is substantial [1]. Clinical manifestations are highly heterogeneous, and patients often undergo prolonged, repeated examinations before receiving a definitive diagnosis [2,3]. Many diseases are associated with recognizable facial phenotypes [4,5]. For example, in families with Crouzon syndrome [6], parents and children show highly overlapping facial features, sharing phenotypes such as mandibular prognathism, shallow orbits, proptosis, and exotropia. Specific facial phenotypes play a key role in the clinical diagnosis of diseases.

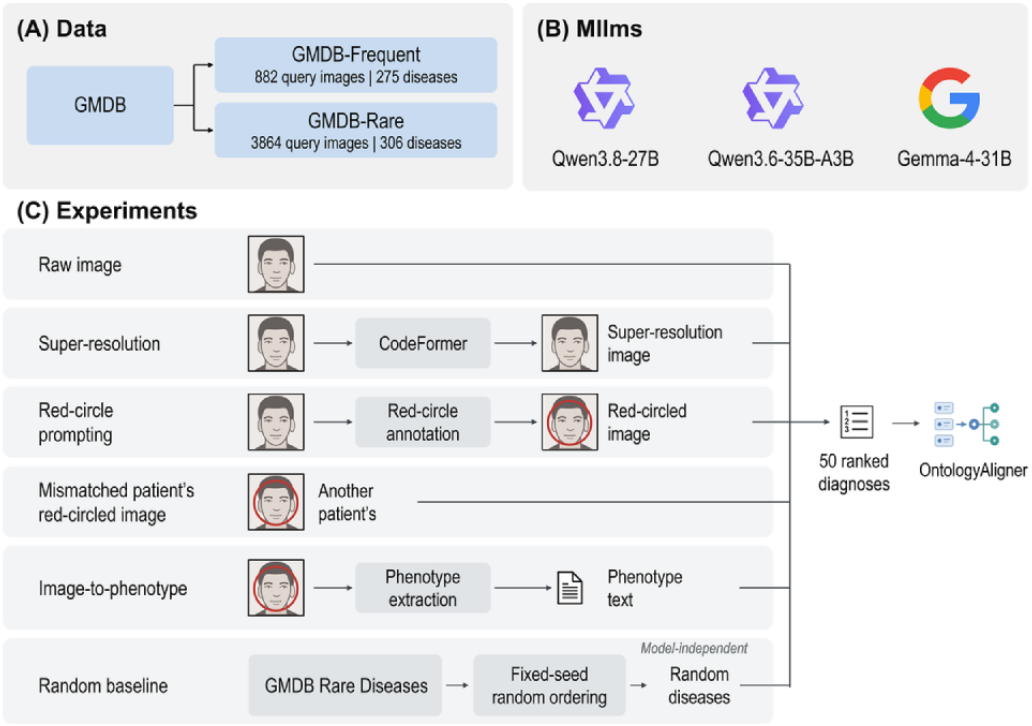
DeepGestalt [7] and GestaltMatcher [8] have shown that facial phenotypes can support the recognition of rare diseases through task-specific training or phenotype-similarity matching [9]. Large language models (LLMs) have also been explored for rare disease care [10,11]. DeepRare, for example, uses an LLM-powered multiagent framework to integrate clinical information, specialized tools, and medical knowledge, generating diagnostic hypotheses with traceable reasoning [12]. However, whether general-purpose multimodal LLMs (MLLMs) [13,14] can effectively prioritize rare genetic disease diagnoses from facial images alone remains an open question. One potential approach to supporting this task is to provide visual prompts that guide the models toward facial regions potentially relevant to disease discrimination [15-18].

Generative image models can produce realistic facial images, suggesting that they capture aspects of facial structure and appearance. This capability may support the localization of facial regions containing potentially relevant phenotypic features. These models can mark such regions in facial images and generate localized visual prompts for MLLMs, whose diagnostic utility can then be evaluated empirically.

To evaluate the facial image–based diagnostic performance of MLLMs and the effects of these prompting strategies, we used the GestaltMatcher Database (GMDB) [19] to compare 3 MLLMs across raw image, super-resolution, red-circle prompting, and 2-stage phenotype-mediated conditions, with mismatched-image controls and a model-independent random baseline. We assessed the clinical relevance and diagnostic utility of visual prompts through physician quality ratings and comparisons with expert annotations. Repeated-inference and uncertainty analyses were used to evaluate the stability of diagnostic accuracy and the relationships between model uncertainty measures and diagnostic hits. Analyses of prediction overlap and disease composition further characterized model outputs. The study workflow is presented in Figure 1. This study aimed to evaluate the diagnostic accuracy of MLLMs for rare genetic diseases using facial images, assess the benefits of image enhancement and visual prompting strategies, and examine the sensitivity of model predictions to patient-specific facial information.

XSL•FO
RenderX

Figure 1. Study workflow. (A) GestaltMatcher Database (GMDB)–Frequent including 882 queries for 275 diseases and GMDB-Rare including 3864 queries for 306 diseases. (B) The 3 multimodal large language models (MLLMs) evaluation of Qwen3.8-27B, Qwen3.6-35B-A3B, and Gemma-4-31B. (C) The diagnostic workflow including raw images, CodeFormer-enhanced images, red-circle prompting, mismatched red-circle controls, and image-to-phenotype text diagnosis, together with a model-independent random baseline. For each query, the models generated 50 ranked candidate diagnoses, which were subsequently mapped to ontology concepts.



Methods

Study Dataset

This study used facial images and corresponding disease labels of patients with rare genetic diseases from the GMDB. On the basis of disease representation in the database, the dataset was divided into 2 subsets: GMDB-Frequent, comprising diseases represented by at least 6 patient images, and GMDB-Rare, comprising diseases represented by fewer than 6 patient images. A 10-fold sampling scheme was applied to GMDB-Rare. GMDB-Frequent included 882 image queries involving 275 diseases, whereas GMDB-Rare included 3864 image queries involving 306 diseases.

Image Processing and Experimental Conditions

CodeFormer (S-Lab, Nanyang Technological University) [20] was used to generate super-resolution images from the raw facial images. We used codeformer-pip v0.0.4 [21] with the official v0.1.0 pretrained weights [22]. The fidelity weight and upscale parameter were set to 0.8 and 2, respectively, with background enhancement and additional face upsampling disabled. We retained the implementation's preprocessing and postprocessing routines: RetinaFace [23] (ResNet-50) detection, OpenCV-based 5-landmark alignment to 512×512 face crops, normalization to [−1, 1], and CodeFormer restoration, followed by inverse mapping and blending into the bilinearly resized image.

Qwen-Image-Edit-2511 [24] was then used to mark potentially abnormal facial regions in these images. Generation used 40 inference steps, true_cfg_scale=4.0, and guidance_scale=1.0. The red circles from the generated images were manually overlaid onto the corresponding CodeFormer super-resolution images to create the final visual prompts. This procedure preserved the facial content of the super-resolution images while incorporating the model-generated regional annotations.

Five input conditions were established for each patient facial image: (1) raw image diagnosis, in which the raw facial image was directly provided to the model; (2) super-resolution diagnosis, in which the CodeFormer-enhanced image was provided to the model; (3) red-circle prompting, in which the super-resolution image with manually overlaid red circles generated by Qwen-Image-Edit-2511 was provided to the model; (4) mismatched red-circle control, in which the current query image was replaced with a red-circle image from another patient in the same subset, with replacement restricted to the same fold in GMDB-Rare and the current query's true diagnosis retained as the reference label; and (5) 2-stage image-to-phenotype diagnosis, in which the model first extracted a facial phenotype description from the current query's red-circle image and then generated candidate diagnoses in a new conversation using only that text as input. All 5 conditions produced 50 candidate diagnoses ranked by likelihood. A random baseline was established by randomly selecting 50 diseases without replacement from the combined pool of 581 distinct diseases across GMDB-Frequent and GMDB-Rare, using Python's random module with a fixed seed of 20260907. The same candidate list was used for all queries in both subsets.

Model Configuration and Diagnostic Inference

This study evaluated 3 MLLMs: Gemma-4-31B [25], Qwen3.6-35B-A3B [26], and Qwen3.8-27B [27]. Within each input condition, all 3 models received the same diagnostic

XSL•FO
RenderX

instructions and generated 50 candidate disease names ranked by likelihood, returning 3 fields in JSON format: ranked_diagnoses, confidence, and insufficient_evidence. The confidence field represented the model's self-reported probability (0%-100%) that the correct diagnosis was included in the complete list of 50 candidates. The insufficient_evidence field indicated whether the model considered the available information insufficient to support a reliable diagnosis. The red-circle prompting, mismatched red-circle, and expert-annotation conditions used the same prompt, with annotation sources masked to the models. The prompts used for facial abnormality annotation with Qwen-Image-Edit-2511 and for the 5 diagnostic conditions are provided in Table S1 in Multimedia Appendix 1.

To align with real-world clinical deployment and meet patient privacy and hospital data governance requirements, all 3 MLLMs and Qwen-Image-Edit-2511 were deployed locally as complete models on hardware comprising an NVIDIA RTX PRO 6000 GPU (NVIDIA Corp) and an Intel Xeon Platinum 8470Q CPU (Intel Corp). All 4 models were run in bfloat16 (BF16) precision without quantization. Details of the deployment environment are provided in Table S2 in Multimedia Appendix 1. The temperature was set to 0.3 for all inference requests, and thinking mode was disabled. Execution errors and invalid outputs, including an incorrect number of diagnoses, triggered retries, with a maximum of 10 retries per request. Nondiagnostic responses, such as abstentions, were retained as experimental outcomes and scored as incorrect for all top-k accuracy metrics. Duplicate diagnoses were retained in their original positions within the 50-item list. After candidate diagnosis generation, OntologyAligner [28] was used to map disease names to Mondo Disease Ontology (MONDO) IDs [29], harmonizing disease synonyms and naming variations and enabling comparisons between predictions and reference labels using standardized disease identifiers. OntologyAligner accomplishes the mapping through a 3-stage process: ontology-aligned retrieval, LLM candidate reranking, and selective hierarchy-guided refinement. OntologyAligner was configured with text-embedding-3-large for embedding, cosine similarity for retrieval, 20 candidate concepts per disease name, and gpt-5.6-sol for reranking at a temperature of 0. Input and output token use is summarized in Table S3 in Multimedia Appendix 1.

To assess mapping accuracy, 500 (2.1%) unique disease names were randomly sampled without replacement from the 23,482 unique names generated across the 3 MLLMs and 5 input conditions. Their MONDO assignments or no-match decisions were manually reviewed, and 97.2% (486/500) were judged correct. To assess sensitivity to simulated MONDO mapping errors, we used the raw image results from all 3 MLLMs in GMDB-Frequent and GMDB-Rare. At each perturbation level from 10% to 90%, in increments of 10 percentage points, we randomly selected the specified proportion of unique disease names. Each selected ID was replaced with a different valid MONDO ID chosen uniformly at random, and the replacement was applied to every occurrence of the same name. We repeated this procedure 500 times per level and recalculated the top-50 accuracy for each model and subset. The 0% level used the original mappings; for each nonzero level, we summarized the mean accuracy and the 2.5th and 97.5th percentiles across simulations.

Visual-Prompt Quality Evaluation

To assess the clinical relevance and localization quality of red-circle prompts, 2 physicians independently reviewed each red-circle image, with access to the corresponding disease label and facial phenotype information. Annotation quality was rated on a 3-level ordinal scale: poor, ordinary, or good. "Good" indicated that the annotated regions accurately highlighted facial abnormalities highly relevant to the target disease and had clear diagnostic value. "Ordinary" indicated that the annotated regions had some relevance but limited specificity or localization accuracy. "Poor" indicated weak relevance to the facial phenotypes associated with the target disease.

To assess the effect of annotation source on diagnostic accuracy, 100 queries were randomly selected from GMDB-Frequent, and 1 physician manually annotated potentially abnormal facial regions on the corresponding super-resolution images. Each of the 3 models performed diagnostic inference using both the expert-annotated images and the images with overlaid red circles generated by Qwen-Image-Edit-2511. Both conditions used identical diagnostic prompts and inference settings, and diagnostic accuracy was compared using a paired analysis of the same queries.

Repeated-Inference Analysis

To assess the stability of diagnostic accuracy across independent inference runs, 100 queries were randomly selected from each of GMDB-Frequent and GMDB-Rare, yielding 200 queries in total. Each of the 3 MLLMs completed 3 independent inference runs on the same queries under the raw image, super-resolution, and red-circle prompting conditions, using the same inference settings as the main experiment. Outputs from each run were saved separately, and disease names were mapped to MONDO IDs using OntologyAligner. Top-50 accuracy and its 95% CI were calculated for each model, input condition, and inference run.

Model Uncertainty Analysis

To assess the relationship between model uncertainty and diagnostic accuracy, we recorded self-reported confidence, evidence-sufficiency judgments, and diagnostic-list perplexity. Confidence represented the estimated probability (0%-100%) that the true diagnosis was included among the 50 candidates, while the insufficient-evidence flag indicated whether the model considered the available information insufficient for a reliable diagnosis. Perplexity was calculated by exponentiating the mean negative log probability of disease-name tokens and was used to quantify the model's uncertainty about the generated sequence, with lower values indicating higher token probabilities:

$$\text{perplexity} = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\ln p_i\right)$$

Here, N is the total number of tokens across all disease names in the candidate diagnosis list; p_i is the probability assigned by the model to the ith disease-name token, conditional on the input and all previously generated content; and ln denotes the natural

XSL•FO
RenderX

logarithm. We then analyzed the associations of self-reported confidence and perplexity with query-level top-50 hit status and compared diagnostic accuracy between queries classified by the models as having sufficient or insufficient evidence.

Prediction-Overlap Analysis

To assess the effects of input conditions on diagnostic predictions, we analyzed the overlap between sets of predicted diseases across conditions and the disease composition of top-1 predictions.

For each query, we calculated the Jaccard index between top-50 predicted disease sets under different input conditions, defined as the size of their intersection divided by the size of their union. Higher values indicate greater overlap between predicted disease sets. Disease-composition analysis compared the distribution of reference labels with that of the top-ranked predicted diseases for each model under each input condition. We separately identified the 5 most frequently occurring diseases and their proportions in the reference labels and each prediction group, grouping all remaining mapped diseases as "other."

Statistical Analysis

Top-50 accuracy was the primary evaluation metric, with top-10, top-20, top-30, and top-40 accuracy also reported. The 95% CIs for accuracy, between-condition differences in accuracy, and other statistics were estimated using 10,000 bootstrap resamples with replacement, with the 2.5th and 97.5th percentiles of the bootstrap distribution serving as the lower and upper bounds. Accuracy was compared between input conditions using 2-sided exact McNemar tests, with *P* values adjusted for multiple comparisons using the Holm method. Uncertainty analyses were restricted to the raw image condition, and Spearman correlation analysis was used to assess the associations of model-reported confidence and diagnostic-list perplexity with top-50 hit status. Absolute between-condition differences in accuracy (percentage points), 95% CIs, and unadjusted and Holm-adjusted *P* values are provided in Multimedia Appendix 2, together with the corresponding correlation results.

Ethical Considerations

This study was not submitted to an ethics committee for review. The study used a legally obtained, publicly available dataset. According to article 32, item 1 of the *Measures for Ethical Review of Life Science and Medical Research Involving Humans* issued by the National Health Commission of the People's Republic of China [30], research using legally obtained public data, or data generated through observation without interfering with public behavior, may be exempted from ethics review. Therefore, under this policy, not applying for ethics board review was appropriate, and no study protocol or application number is applicable.

Results

Diagnostic Performance Across Models and Input Conditions

Figure 2A presents the top-50 diagnostic accuracy for 882 queries in GMDB-Frequent, with accuracy ranging from 13.95% to 18.25% across the 3 models under 5 input conditions. Figure 2B presents the results for 3864 queries in GMDB-Rare, with corresponding accuracy ranging from 0.57% to 1.35%.

In comparisons of image processing and visual prompting conditions, the red-circle condition yielded numerically higher accuracy than the raw image condition in 5 of the 6 model-subset combinations and equal accuracy in 1. In GMDB-Frequent, accuracy for Gemma-4-31B, Qwen3.6-35B-A3B, and Qwen3.8-27B changed from 15.87%, 14.51%, and 15.99% to 16.89%, 15.19%, and 16.44%, respectively, with $P>.99$ for all 3 comparisons. In GMDB-Rare, the corresponding accuracy changed from 0.91%, 0.91%, and 0.57% to 0.93%, 0.91%, and 1.27%, respectively. Qwen3.8-27B showed an increase of 0.70 percentage points (adjusted $P=.009$), while the other 2 models showed no significant changes ($P>.99$). The direction of accuracy changes for super-resolution vs raw images and for red-circle images vs super resolution varied by model and subset, with $P>.05$ for all comparisons.

In comparisons of patient-image matching, accuracy under the matched red-circle condition was numerically higher than under the mismatched red-circle condition across all 3 models and both subsets. The advantage ranged from 0.68 to 1.25 percentage points in GMDB-Frequent and from 0.26 to 0.49 percentage points in GMDB-Rare. All 6 comparisons yielded $P>.05$.

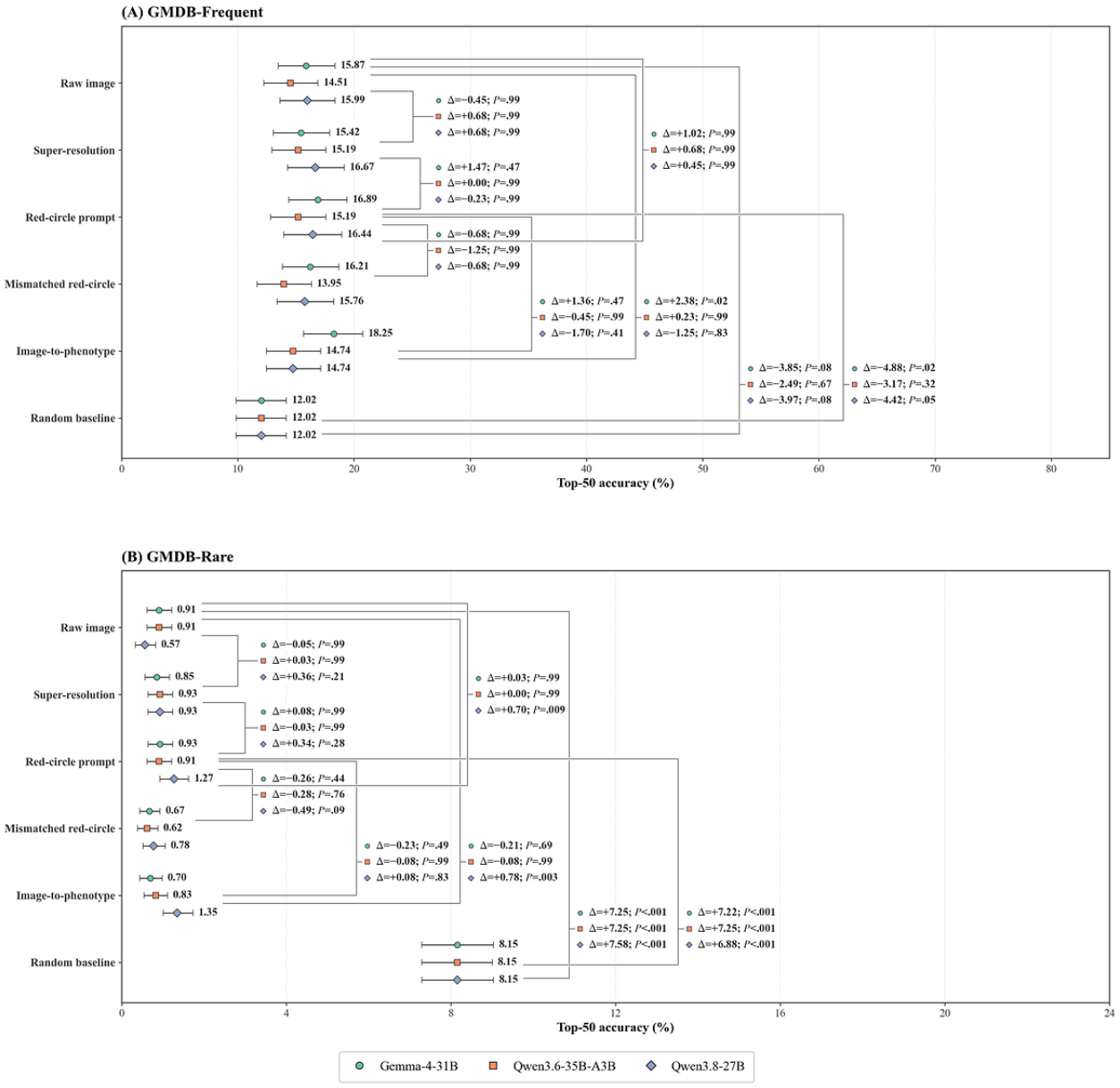
In comparisons of the 2-stage diagnostic baseline, image-to-phenotype yielded significant improvements over the raw image condition in 2 combinations: Gemma-4-31B in GMDB-Frequent, with accuracy increasing from 15.87% to 18.25% (an increase of 2.38 percentage points; adjusted $P=.02$), and Qwen3.8-27B in GMDB-Rare, with accuracy increasing from 0.57% to 1.35% (an increase of 0.78 percentage points; adjusted $P=.003$). Comparisons against the raw image condition in the remaining 4 combinations and against the red-circle condition in all 6 combinations yielded $P>.05$.

The random baseline achieved an accuracy of 12.02% in GMDB-Frequent and 8.15% in GMDB-Rare. In GMDB-Frequent, accuracy under both the raw image and red-circle conditions was numerically higher than the random baseline for all 3 models. For Gemma-4-31B, top-50 accuracy was significantly higher under the red-circle condition than under the random baseline (16.89% vs 12.02%, respectively; adjusted $P=.02$), while the other 5 comparisons yielded $P>.05$. In GMDB-Rare, the random baseline significantly outperformed all 3 models under both the raw image and red-circle conditions (all $P<.001$). Accuracy at different candidate-list cutoffs ($k=10$, 20, 30, 40, and 50) is presented in Figure S1 in Multimedia Appendix 1, and complete statistical results are provided in Multimedia Appendix 2.

Figure S2 in Multimedia Appendix 1 presents the sensitivity of raw-image top-50 accuracy to simulated MONDO mapping errors. Mean accuracy declined as the proportion of reassigned disease names increased across all 3 models and both GMDB subsets. At 10% reassignment, the mean decreases in accuracy across 500 simulations were 1.52 (SD 0.92), 1.34 (SD 1.04), and 1.50 (SD 1.30) percentage points in GMDB-Frequent and 0.08 (SD 0.14), 0.08 (SD 0.10), and 0.04 (SD 0.08) percentage points in GMDB-Rare for Gemma-4-31B, Qwen3.6-35B-A3B, and Qwen3.8-27B, respectively.

XSL•FO
RenderX

Figure 2. Top-50 accuracy under 5 model input conditions and the random baseline. (A) GestaltMatcher Database (GMDB)–Frequent (n=882) and (B) GMDB-Rare (n=3864). Points and error bars indicate accuracy and 95% bootstrap CIs, respectively. Differences are calculated as the accuracy of the lower condition minus that of the upper condition and are expressed in percentage points. *P* values are from 2-sided exact McNemar tests with Holm adjustment.



Visual-Prompt Quality and Diagnostic Benefit

Figure 3A presents the distribution of ratings independently assigned by 2 physicians to 4746 red-circle annotations generated by Qwen-Image-Edit-2511, categorized as poor, ordinary, or good. Ordinary was the predominant rating for both physicians, accounting for 48.9% (431/882) to 50.8% (448/882) of annotations in GMDB-Frequent and 61.6% (2381/3864) to 63.1% (2437/3864) in GMDB-Rare. The proportion rated good ranged from 31.5% (278/882) to 35.6% (314/882) in GMDB-Frequent and from 16.1% (624/3864) to 17.7% (683/3864) in GMDB-Rare. Overall agreement between the 2 physicians was 71.8% (3409/4746).

Figure 3B presents the quality-stratified analyses of GMDB-Frequent conducted separately using each physician's ratings. In the good subgroup, Gemma-4-31B achieved significantly higher top-50 accuracy under the red-circle condition than under the super-resolution condition. In evaluation 1, accuracy increased from 17.27% to 23.02%, a gain of 5.76 percentage points (adjusted *P*=.005); in evaluation 2, accuracy increased from 17.20% to 21.66%, a gain of 4.46 percentage points (adjusted *P*=.048). All remaining stratified comparisons in GMDB-Frequent yielded adjusted *P*>.05.

Figure 3C presents the quality-stratified results for GMDB-Rare, where comparisons for all models across all quality subgroups yielded adjusted *P*>.05. Significant improvements in the quality-stratified analyses were confined to Gemma-4-31B in

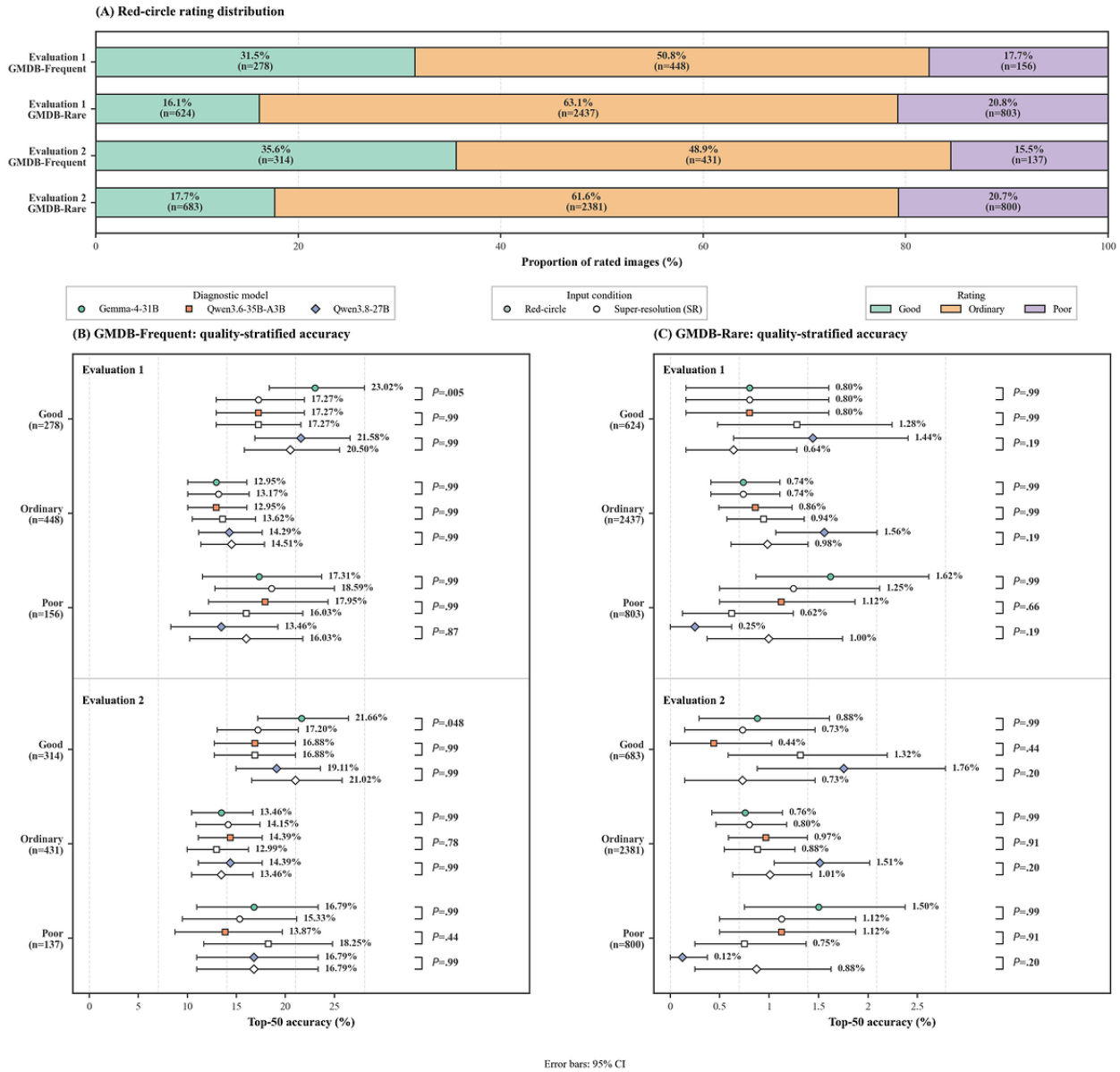
XSL•FO
RenderX

the GMDB-Frequent good subgroups defined separately by the 2 physicians' ratings.

Under the red-circle condition in GMDB-Frequent, the good subgroup generally showed numerically higher accuracy. For Gemma-4-31B and Qwen3.8-27B in both evaluations and for Qwen3.6-35B-A3B in evaluation 2, accuracy in the good subgroup exceeded that in the corresponding ordinary and poor subgroups.

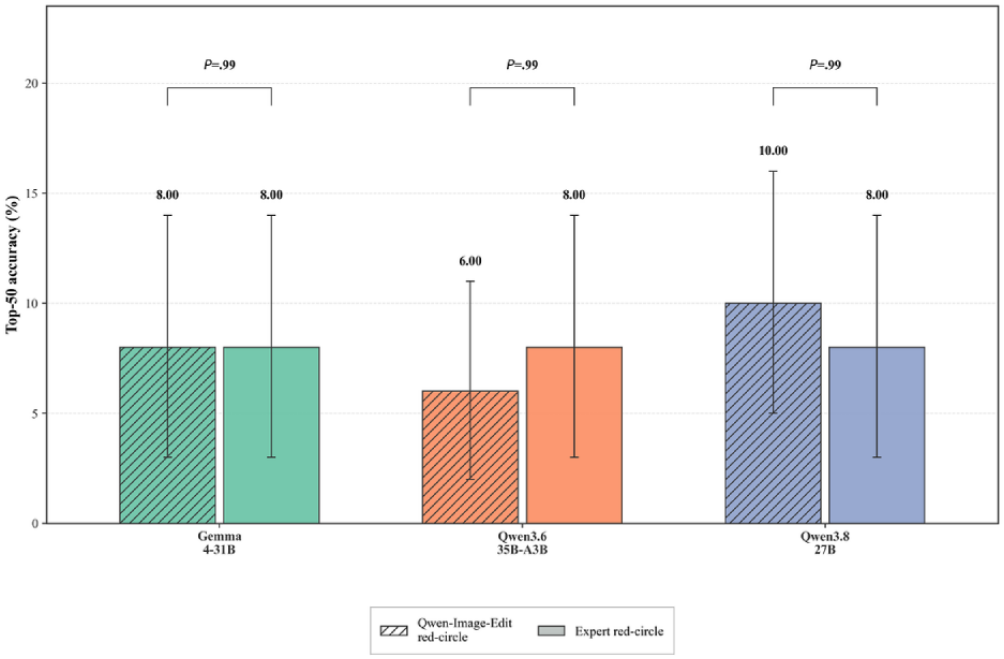
Figure 4 compares diagnostic accuracy between expert-annotated and model-generated circle images. A physician manually circled abnormal facial regions on 100 super-resolution images from GMDB-Frequent. Each of the 3 models received both expert-annotated and Qwen-Image-Edit-2511–generated red-circle overlays for the same set of queries. Under the expert-annotation condition, top-50 accuracy was 8.00% for Gemma-4-31B, Qwen3.6-35B-A3B, and Qwen3.8-27B. Under the model-generated annotation condition, the corresponding accuracy was 8.00%, 6.00%, and 10.00%, respectively. Compared with model-generated annotations, expert annotations yielded equal accuracy for Gemma-4-31B, an increase of 2.00 percentage points for Qwen3.6-35B-A3B, and a decrease of 2.00 percentage points for Qwen3.8-27B. All 3 paired comparisons yielded adjusted $P>.99$. Detailed statistical results are provided in Multimedia Appendix 2.

Figure 3. (A) Proportions and sample sizes of the good, ordinary, and poor categories based on independent ratings by 2 physicians. (B) Top-50 accuracy stratified by red-circle annotation quality in GestaltMatcher Database (GMDB)–Frequent. (C) Top-50 accuracy stratified by red-circle annotation quality in GMDB-Rare. Filled circles indicate red-circle inputs, open circles indicate super-resolution inputs, and error bars represent 95% bootstrap CIs. *P* values were adjusted using the Holm method.



XSL•FO
RenderX

Figure 4. Comparison of top-50 diagnostic accuracy between expert-annotated and Qwen-Image-Edit-2511–generated circle images. All 3 models received both types of red-circle annotations for the same 100 GestaltMatcher Database (GMDB)–Frequent queries using identical diagnostic prompts. Error bars represent 95% bootstrap CIs, and brackets indicate Holm-adjusted *P* values from 2-sided exact McNemar tests.



Stability of Repeated Diagnostic Inference

To assess the stability of diagnostic performance across runs, we performed 3 independent inference runs on a fixed set of 200 queries, comprising 100 queries each from GMDB-Frequent and GMDB-Rare. Each model was evaluated under the raw image, super-resolution, and red-circle conditions using the same settings as in the main experiment.

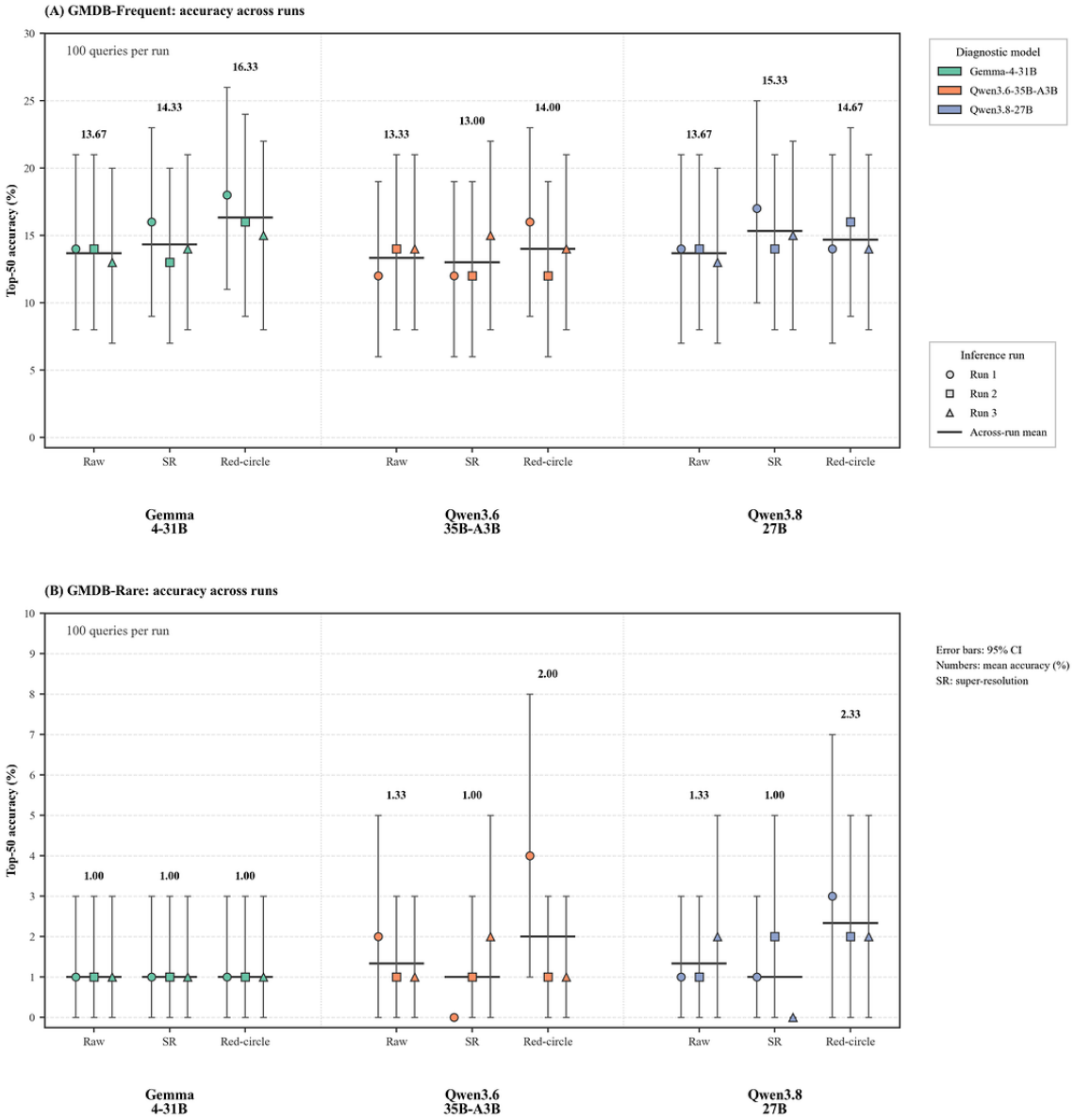
Figure 5A presents the results for GMDB-Frequent. Across model-input combinations, the difference between the maximum and minimum top-50 accuracy over the 3 runs ranged from 1 to 4 percentage points. Across the 3 runs, mean accuracy for the raw image, super-resolution, and red-circle conditions was 13.67%, 14.33%, and 16.33%, respectively, for Gemma-4-31B; 13.33%, 13.00%, and 14.00% for Qwen3.6-35B-A3B; and 13.67%, 15.33%, and 14.67% for Qwen3.8-27B. Gemma-4-31B and Qwen3.6-35B-A3B achieved their highest mean accuracy under the red-circle condition, while Qwen3.8-27B achieved its highest mean accuracy under the super-resolution condition.

Figure 5B presents the results for GMDB-Rare. Across model-input combinations, the difference between the maximum and minimum accuracy over the 3 runs ranged from 0 to 3 percentage points. Gemma-4-31B achieved a mean accuracy of 1.00% under all 3 conditions. Across the raw image, super-resolution, and red-circle conditions, mean accuracy was 1.33%, 1.00%, and 2.00%, respectively, for Qwen3.6-35B-A3B and 1.33%, 1.00%, and 2.33%, respectively, for Qwen3.8-27B.

Across these 200 fixed queries, top-50 accuracy showed limited variation over the 3 independent inference runs. The low number of successful hits in GMDB-Rare means that changes in hit status for a small number of queries can affect accuracy.

XSL•FO
RenderX

Figure 5. Variation in top-50 accuracy across 3 independent inference runs. (A) GestaltMatcher Database (GMDB)–Frequent and (B) GMDB-Rare. Each model completed 3 independent inference runs under each input condition using the same fixed set of 100 queries per subset. Error bars represent 95% bootstrap CIs for accuracy in each run. Horizontal lines and the values above them indicate the arithmetic mean accuracy across the 3 runs. SR: super resolution.



Model Uncertainty and Diagnostic Accuracy

To assess the relationship between uncertainty metrics and diagnostic accuracy, we asked the models to report diagnostic confidence and evidence sufficiency. We also recorded output-token log probabilities and calculated perplexity.

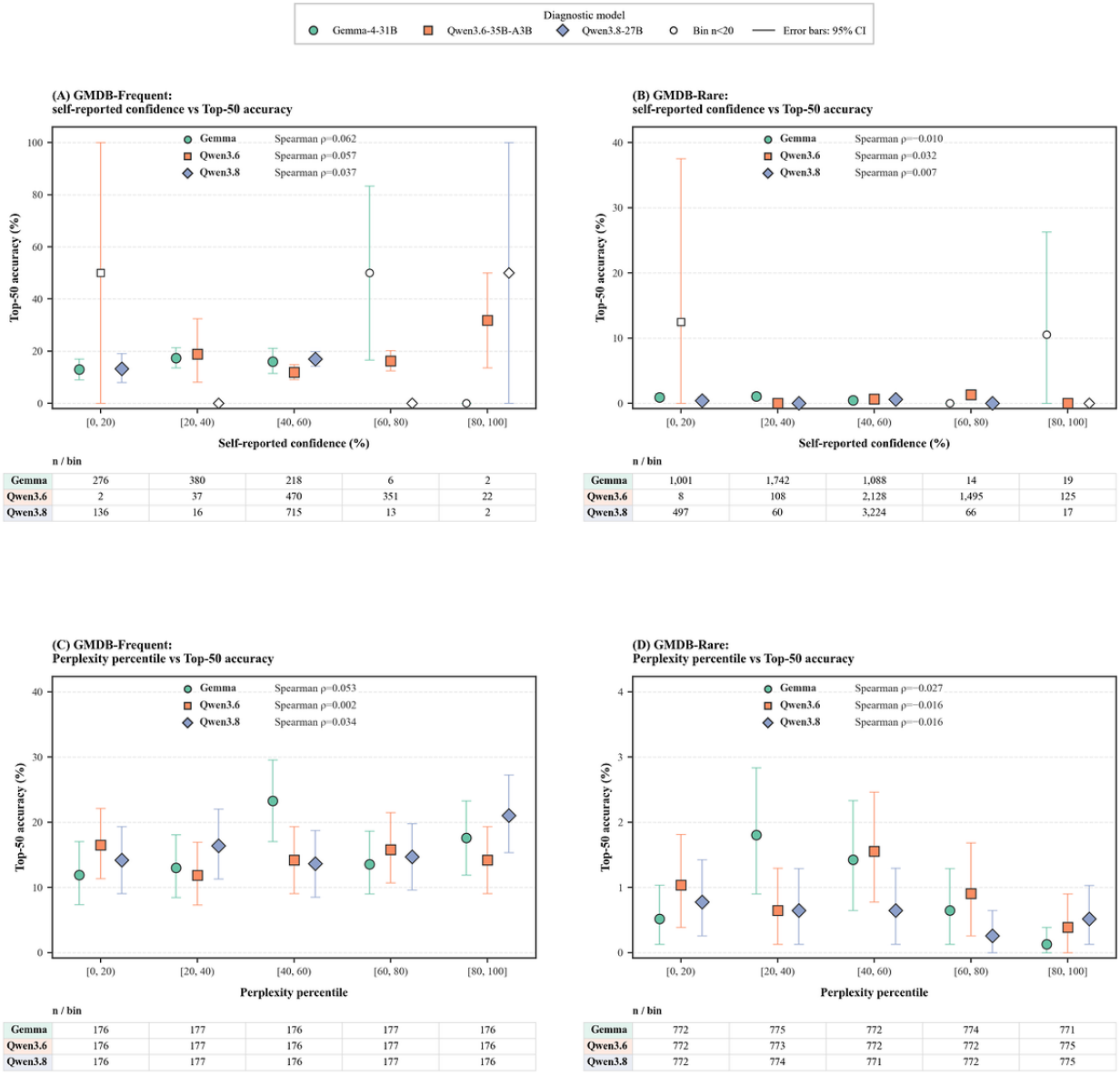
Figure 6A shows that the Spearman correlation coefficients between self-reported confidence and top-50 accuracy in GMDB-Frequent were 0.062, 0.057, and 0.037. Figure 6B shows the corresponding coefficients in GMDB-Rare, which were −0.010, 0.032, and 0.007.

Figure 6C shows that the correlation coefficients between perplexity and top-50 accuracy in GMDB-Frequent were 0.053, 0.002, and 0.034. Figure 6D shows the corresponding coefficients in GMDB-Rare, which were −0.027, −0.016, and −0.016. All correlation coefficients had absolute values below 0.10 and all Holm-adjusted P>.05, indicating weak associations of self-reported confidence and perplexity with diagnostic hits.

Figure S3 in Multimedia Appendix 1 presents top-50 accuracy stratified by the models' assessments of evidence sufficiency. Under the raw image condition, all 3 models predominantly reported "insufficient evidence," and the direction of between-group accuracy differences varied by model and subset. Detailed statistical results are provided in Multimedia Appendix 2.

XSL•FO
RenderX

Figure 6. Relationships between uncertainty metrics and top-50 diagnostic accuracy under the raw image condition. (A) Top-50 accuracy stratified by model self-reported confidence in GestaltMatcher Database (GMDB)–Frequent. (B) Top-50 accuracy stratified by model self-reported confidence in GMDB-Rare. (C) Top-50 accuracy stratified by model-specific perplexity percentiles in GMDB-Frequent. (D) Top-50 accuracy stratified by model-specific perplexity percentiles in GMDB-Rare. Points and error bars indicate within-group accuracy and its 95% bootstrap CI, respectively. Open circles indicate groups with fewer than 20 queries. Spearman ρ was calculated using query-level uncertainty metrics and top-50 hit status.



Prediction Overlap and Composition

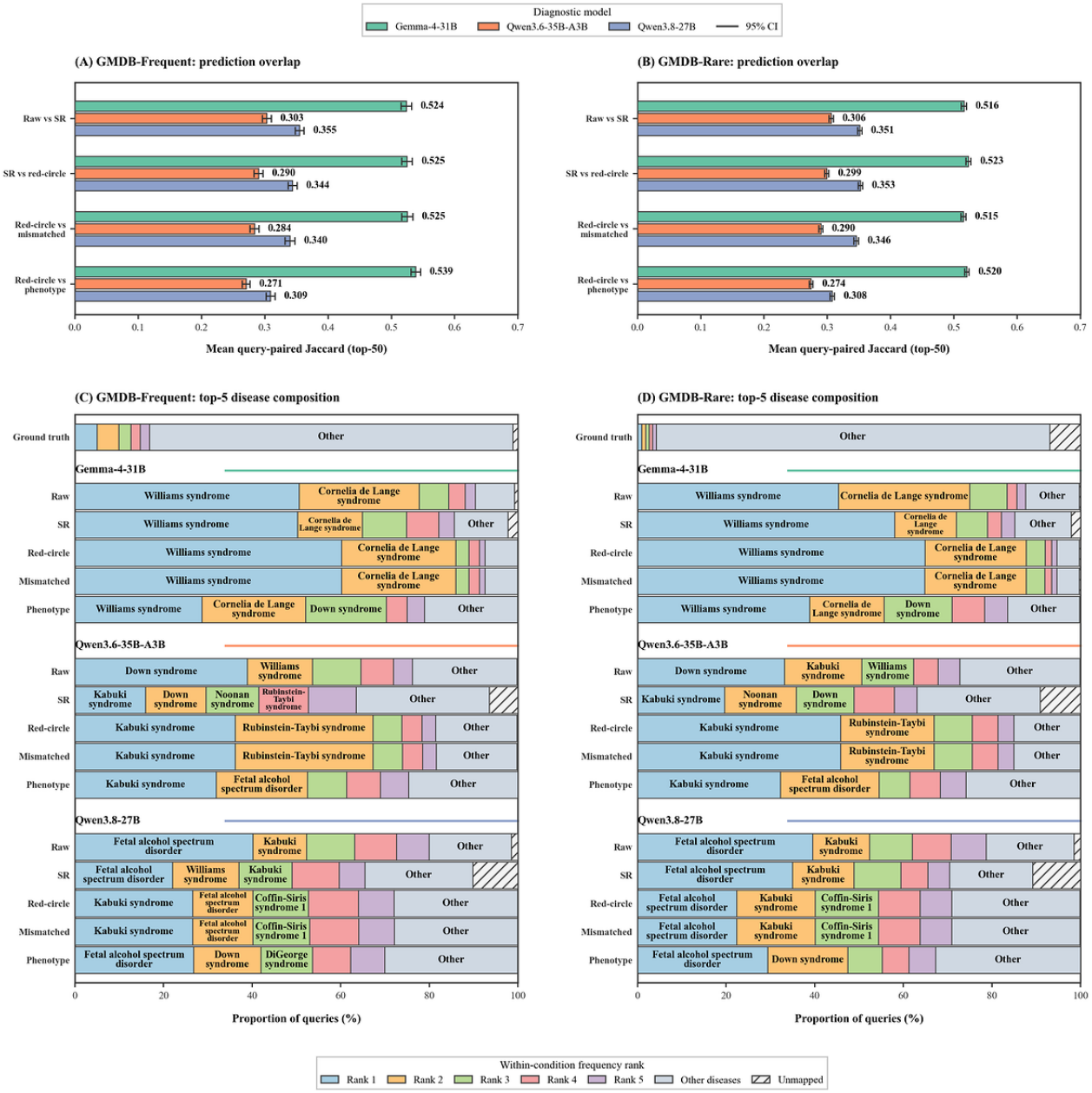
The top-50 candidate diagnosis sets generated by each model showed substantial overlap across input conditions (Figure 7A for GMDB-Frequent and Figure 7B for GMDB-Rare). Gemma-4-31B had the highest mean pairwise Jaccard index, approximately 0.52 in both subsets; the corresponding values were approximately 0.27 to 0.31 for Qwen3.6-35B-A3B and 0.31 to 0.36 for Qwen3.8-27B. The overlap between candidate sets under the red-circle and mismatched red-circle conditions was similar to that observed across different processing conditions for the same image. This finding suggests limited changes in candidate sets when the patient image was replaced.

Top-1 predictions from all 3 models were highly concentrated in a small number of diseases. In GMDB-Frequent (Figure 7C), the 5 most frequent diseases in the ground-truth labels accounted for 16.8% (148/882) of all queries. Under the red-circle condition, the 5 most frequently predicted diseases accounted for 92.5% (816/882), 81.4% (718/882), and 72% (635/882) of predictions from Gemma-4-31B, Qwen3.6-35B-A3B, and Qwen3.8-27B, respectively. In GMDB-Rare (Figure 7D), the corresponding proportion for the ground-truth labels was 4.2% (162/3864), while the proportions for the 3 models reached 94.7% (3658/3864), 85% (3283/3864), and 71% (2740/3864), respectively.

XSL•FO
RenderX

Gemma-4-31B's predictions were primarily concentrated in Williams syndrome and Cornelia de Lange syndrome, while Qwen3.6-35B-A3B's predictions were primarily concentrated in Kabuki syndrome and Rubinstein-Taybi syndrome. Qwen3.8-27B frequently predicted Kabuki syndrome, fetal alcohol spectrum disorder, and Coffin-Siris syndrome 1. Similar concentration patterns were observed under the mismatched red-circle condition and the 2-stage image-to-phenotype baseline.

Figure 7. Prediction overlap and top-1 disease composition. (A) Overlap of top-50 candidate diagnosis sets across input conditions in GestaltMatcher Database (GMDB)–Frequent. (B) Overlap of top-50 candidate diagnosis sets across input conditions in GMDB-Rare. (C) Disease composition of the ground-truth labels and each model's top-1 predictions under different input conditions in GMDB-Frequent. (D) Disease composition of the ground-truth labels and each model's top-1 predictions under different input conditions in GMDB-Rare. SR: super resolution.



Discussion

Principal Findings

This study evaluated the ability of 3 locally deployed MLLMs to perform diagnosis based on facial images and examined the effects of image enhancement, red-circle prompting, and the 2-stage image-to-phenotype baseline. Overall diagnostic accuracy was limited, with particularly low performance on GMDB-Rare. Compared with the raw image condition, red-circle prompting produced numerical increases across most model-subset combinations, with a statistically significant improvement observed for Qwen3.8-27B on GMDB-Rare. The quality-stratified analysis further showed that Gemma-4-31B benefited from red-circle prompting in the good subgroup of GMDB-Frequent. The 2-stage image-to-phenotype baseline also improved accuracy for specific models and subsets. Matching red circles yielded numerical accuracy increases

XSL•FO
RenderX

relative to mismatched red circles across all comparisons, although statistical support for these differences was limited. Together with the overlap among prediction sets after patient-image replacement and the concentration of top-1 outputs, these findings suggest that model outputs were only weakly constrained by patient-specific facial information. Improvements associated with different input strategies were concentrated in specific models and disease subsets, and their applicability requires further validation. In addition, overall accuracy varied only slightly across repeated inferences, while self-reported confidence and diagnostic-list perplexity showed weak associations with diagnostic accuracy.

Diagnostic Utility of Visual and Phenotype-Based Prompting

Comparisons between the red-circle and raw image conditions reflect the combined effects of image enhancement and circle annotation. Comparisons between the red-circle and super-resolution conditions provide a more direct assessment of regional prompting. Gemma-4-31B showed a significant benefit in the good subgroup of GMDB-Frequent, suggesting that clinically relevant regional localization may help the model use facial cues. The higher accuracy observed in the good subgroup may also relate to case composition and phenotype recognizability. In the expert-annotated circle experiment, the direction of the difference between the 2 circle sources varied across models, with limited statistical support. These findings suggest that circle quality and the model's ability to use local features for disease discrimination should be evaluated together.

The 2-stage image-to-phenotype baseline converts visual cues into explicit phenotype descriptions, which may help the model organize features and retrieve relevant disease knowledge. Its significant improvements relative to the raw image condition were concentrated in specific models and subsets. Comparisons with the red-circle condition yielded Holm-adjusted $P>.05$ across all settings, and the additional diagnostic value of phenotype text requires further validation.

Stability and Uncertainty

Overall accuracy varied only slightly across the 3 repeated-inference runs, indicating a degree of run-to-run stability in diagnostic performance on the selected queries. This finding reflects variation in the overall hit rate across runs; prediction consistency for individual queries and the stability of the benefits of visual prompting require further evaluation. GMDB-Rare had a low number of diagnostic hits, so changes in the outcomes of a small number of queries could affect accuracy. Its stability should therefore be interpreted in the context of sample size.

Self-reported confidence and perplexity showed weak associations with diagnostic hits, and the direction of accuracy differences between evidence-sufficiency groups varied across models and subsets. These findings suggest that these metrics provide limited information about diagnostic reliability in the current task. Self-reported confidence reflects the model's subjective estimate that the candidate list contains the correct diagnosis, while perplexity reflects the token probabilities of the generated sequence of disease names. The use of self-reported confidence, perplexity, or evidence-sufficiency judgments to identify diagnostic outputs with low reliability therefore requires further validation [31].

Limitations

This study has several limitations. First, there is a potential risk of training-data contamination [32,33]: publicly available facial images from the evaluated cases, such as patient photographs published in case reports, may have been included in model training data, potentially affecting estimates of diagnostic accuracy and assessments of generalizability. Limited disclosure of case-level sources in model training data makes this overlap difficult to quantify directly. Second, clinical comparisons were limited to physician ratings of circle quality and a comparison between expert-annotated and automatically generated circles for 100 queries. These analyses primarily assessed the clinical relevance of the annotations and their impact on model diagnostic performance. Direct comparisons with physician diagnoses and the clinical utility of the models as assistive tools require further evaluation. Third, real-world population prevalence data for individual diseases are limited, and a disease-frequency baseline remains to be added. In preliminary image-free controls, all models declined to answer because images were unavailable. The respective contributions of disease priors and visual information to diagnostic ranking require further clarification. Fourth, repeated inference covered only 200 queries and 3 runs, reflecting run-to-run variability within the selected subset. The small number of diagnostic hits in GMDB-Rare also made the results sensitive to changes in the outcomes of a small number of queries. Fifth, simulated MONDO ID substitutions provide an indirect assessment of sensitivity to mapping errors. However, they may differ from actual mapping errors, so this analysis cannot quantify the effect of actual mapping errors on diagnostic accuracy. Sixth, diagnostic accuracy depends on the normalization of disease names to MONDO. Manual verification of 500 unique disease names found that 486 (97.2%) mappings were correct. Residual mapping errors may affect accuracy estimates and comparisons across conditions. Seventh, the disease spectrum, age distribution, ancestry, and image acquisition conditions of the disease-labeled facial images in GMDB may differ from those encountered in routine outpatient practice. The effects of these differences on model performance remain unquantified in this study. Previous studies on facial image–assisted diagnosis have also highlighted the influence of image quality and population coverage on system performance [34]. Eighth, the information used in this study was limited to facial images and model-generated phenotype descriptions. Performance incorporating comprehensive physician-verified phenotypes, developmental history, family history, and genetic testing information requires further evaluation. Ninth, the absence of query-level inference runtime records limited the assessment of runtime performance.

Future Directions

Future research could compare visual prompts based on points, bounding boxes, and masks, as well as in-context examples, to evaluate the effects of different guidance strategies. Facial landmark detection could provide standardized geometric

XSL•FO
RenderX

localization, while disease-related phenotype annotations could support the training and validation of abnormality-recognition modules to generate regional prompts and assess their localization quality and diagnostic utility. For information integration, facial analysis could be combined with human phenotype ontology terms [35], family history, and genetic testing results [36,37]. Multimodal retrieval-augmented generation could further incorporate relevant cases and disease knowledge [38-40] to evaluate the contributions of more comprehensive patient information and external evidence to diagnosis.

To investigate training-data exposure, future research could explore the number of disease-related publications in PubMed as a proxy for literature exposure and analyze its association with top-50 accuracy. Manual verification of MONDO mappings could also be expanded to identify disease names prone to mapping errors and common error patterns, guide improvements to the mapping tool, and assess the impact of mapping corrections on diagnostic accuracy. To assess run-to-run stability, the number of queries and repeated-inference runs could be increased to further quantify fluctuations in overall accuracy across runs. In addition, the relationships of self-reported confidence, perplexity, and evidence sufficiency with disease frequency, image quality, and prompt quality should be analyzed to evaluate the ability of these metrics to identify diagnostic outputs with low reliability.

Clinical validation should compare diagnoses made by physicians and models under the same information conditions and evaluate accuracy, diagnostic time, and error types for physicians making independent and model-assisted judgments. Ideally, further studies should incorporate prospective cases from multiple centers and conduct stratified analyses by age, ethnicity, and disease spectrum to clarify the scope of clinical applicability and model performance across different populations.

Conclusions

Under the experimental settings evaluated in this study, Gemma-4-31B, Qwen3.6-35B-A3B, and Qwen3.8-27B showed limited ability to diagnose rare genetic diseases using facial images alone. Performance was particularly low on GMDB-Rare. The benefits of image enhancement, red-circle prompting, and the 2-stage image-to-phenotype workflow were primarily observed in specific models and disease subsets. Together, the mismatched-image controls, prediction-set overlap, and disease-composition analyses suggest that model outputs were only weakly constrained by patient-specific facial information. Future research should strengthen the integration of patient phenotypes with disease knowledge and validate the models' clinical utility as assistive tools through direct comparisons with physicians and multicenter prospective studies.

Funding

This work was supported by the National Natural Science Foundation of China (grants 32570773 and 32270690) and the Pioneer and Leading Goose R&D Program of Zhejiang Province (grant 2026C01022). The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Data Availability

The GestaltMatcher Database [41] is available online. The code, prompts, random seeds, and model outputs are publicly available at GitHub [42]. The version corresponding to this manuscript is identified by the following Git commit hash: 369cbfff2c28094b77d7815c866eb3fd3f094e25.

Generative AI Use

The authors declare that generative AI was used in the creation of this manuscript. Generative AI (Deepseek) was used for proofreading, editing, or language polishing. The authors verified all outputs of the tool(s) used in the preparation of this manuscript and take full responsibility for the final content of the manuscript.

Authors' Contributions

JS had full access to the data and takes responsibility for the integrity of the data and the accuracy of the data analysis. JS contributed to conceptualization, methodology, validation, visualization, software, writing—original draft. ZX contributed to writing—review and editing. MX contributed to writing—review and editing. CB contributed to writing—review and editing. YZ contributed to writing—review and editing. XZ contributed to writing—review and editing. XL contributed to writing—review and editing. ZL contributed to validation. QC contributed to validation. BS contributed to supervision, project administration, funding acquisition, and writing—review and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Model prompts, computing environment, and supplementary analyses.

XSL•FO
RenderX

[DOCX File , 1475 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Detailed statistical comparisons.
[XLSX File (Microsoft Excel File), 27 KB-Multimedia Appendix 2]

References

1. Nguengang Wakap S, Lambert DM, Olry A, Rodwell C, Gueydan C, Lanneau V, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. Feb 2020;28(2):165-173. [FREE Full text] [doi: 10.1038/s41431-019-0508-0] [Medline: 31527858]
2. Bauskis A, Strange C, Molster C, Fisher C. The diagnostic odyssey: insights from parents of children living with an undiagnosed condition. Orphanet J Rare Dis. Jun 18, 2022;17(1):233. [FREE Full text] [doi: 10.1186/s13023-022-02358-x] [Medline: 35717227]
3. Faye F, Crocione C, Anido de Peña R, Bellagambi S, Escati Peñaloza L, Hunter A, et al. Time to diagnosis and determinants of diagnostic delays of people living with a rare disease: results of a Rare Barometer retrospective patient survey. Eur J Hum Genet. Sep 2024;32(9):1116-1126. [doi: 10.1038/s41431-024-01604-z] [Medline: 38755315]
4. Wenderott K, Krups J, Zaruchas F, Krawitz P, Weigl M, Lesmann H. Workflow analysis and evaluation of a next-generation phenotyping tool: a qualitative study of Face2Gene. Eur J Hum Genet. Oct 2025;33(10):1333-1341. [doi: 10.1038/s41431-025-01875-0] [Medline: 40410386]
5. Hart TC, Hart PS. Genetic studies of craniofacial anomalies: clinical implications and applications. Orthod Craniofac Res. Aug 2009;12(3):212-220. [FREE Full text] [doi: 10.1111/j.1601-6343.2009.01455.x] [Medline: 19627523]
6. Li ZL, Chen X, Zhuang WJ, Zhao W, Liu YN, Zhang FX, et al. FGFR2 mutation in a Chinese family with unusual Crouzon syndrome. Int J Ophthalmol. Oct 18, 2016;9(10):1403-1408. [FREE Full text] [doi: 10.18240/ijo.2016.10.06] [Medline: 27803855]
7. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. Jan 2019;25(1):60-64. [doi: 10.1038/s41591-018-0279-0] [Medline: 30617323]
8. Hsieh TC, Bar-Haim A, Moosa S, Ehmke N, Gripp KW, Pantel JT, et al. GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. Nat Genet. Mar 2022;54(3):349-357. [FREE Full text] [doi: 10.1038/s41588-021-01010-x] [Medline: 35145301]
9. Song J, He M, Zheng X, Zhang Y, Bi C, Feng J, et al. Face-based machine learning diagnostics: applications, challenges and opportunities. Artif Intell Rev. May 13, 2025;58:243. [doi: 10.1007/s10462-025-11246-2]
10. Yang J, Shu L, Duan H, Li H. RDguru: a conversational intelligent agent for rare diseases. IEEE J Biomed Health Inform. Sep 2025;29(9):6366-6378. [doi: 10.1109/JBHI.2024.3464555] [Medline: 39298307]
11. Song J, Xu Z, Xiao M, Bi C, Zhang Y, Zheng X, et al. Initial-visit specialty triage in rare diseases using large language models: retrospective benchmarking study. J Med Internet Res. Jul 23, 2026;28:e101711. [FREE Full text] [doi: 10.2196/101711] [Medline: 42490550]
12. Zhao W, Wu C, Fan Y, Qiu P, Zhang X, Sun Y, et al. An agentic system for rare disease diagnosis with traceable reasoning. Nature. Mar 2026;651(8106):775-784. [doi: 10.1038/s41586-025-10097-9] [Medline: 41708847]
13. Bradshaw TJ, Tie X, Warner J, Hu J, Li Q, Li X. Large language models and large multimodal models in medical imaging: a primer for physicians. J Nucl Med. Feb 03, 2025;66(2):173-182. [FREE Full text] [doi: 10.2967/jnumed.124.268072] [Medline: 39819692]
14. Chakraborty C, Bhattacharya M, Pal S, Chatterjee S, Das A, Lee SS. AI-enabled language models (LMs) to large language models (LLMs) and multimodal large language models (MLLMs) in drug discovery and development. J Adv Res. Dec 2025;78:377-389. [FREE Full text] [doi: 10.1016/j.jare.2025.02.011] [Medline: 39952319]
15. Chen X, Lai Z, Ruan K, Chen S, Liu J, Liu Z. R-LLaVA: improving med-VQA understanding through visual region of interest. In: 2025 International Joint Conference on Neural Networks (IJCNN). New York, NY. IEEE; 2025.
16. Wu M, Cai X, Ji J, Li J, Huang O, Luo G, et al. ControlMLLM: training-free visual prompt learning for multimodal large language models. In: NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems. New York, NY. Curran Associates Inc; 2024:45206-45234.
17. Sun Z, Fang Y, Wu T, Zhang P, Zang Y, Kong S, et al. Alpha-CLIP: a CLIP model focusing on wherever you want. arXiv. Preprint posted online on December 6, 2023. [doi: 10.48550/arXiv.2312.03818]
18. Schramm S, Preis S, Metz MC, Jung K, Schmitz-Koep B, Zimmer C, et al. Impact of multimodal prompt elements on diagnostic performance of GPT-4V in challenging brain MRI cases. Radiology. Jan 2025;314(1):e240689. [doi: 10.1148/radiol.240689] [Medline: 39835982]
19. Lesmann H, Hustinx A, Moosa S, Klinkhammer H, Marchi E, Caro P, et al. GestaltMatcher database - a global reference for facial phenotypic variability in rare human diseases. medRxiv. Preprint posted online on October 8, 2024. [FREE Full text] [doi: 10.1101/2023.06.06.23290887] [Medline: 37503210]

XSL•FO
RenderX

20. Zhou S, Chan KC, Li C, Loy CC. Towards robust blind face restoration with codebook lookup transformer. arXiv. Preprint posted online on June 22, 2022. [doi: 10.48550/arXiv.2206.11253]
21. codeformer-pip 0.0.4. PyPI. URL: https://pypi.org/project/codeformer-pip/0.0.4/ [accessed 2026-10-21]
22. CodeFormer V0.1.0 release. GitHub. URL: https://github.com/sczhou/CodeFormer/releases/tag/v0.1.0 [accessed 2026-10-21]
23. Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S. RetinaFace: single-shot multi-level face localisation in the wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New York, NY. IEEE; 2020.
24. Qwen/Qwen-Image-Edit-2511. Hugging Face. URL: https://huggingface.co/Qwen/Qwen-Image-Edit-2511 [accessed 2026-10-21]
25. google/gemma-4-31B-it. Hugging Face. URL: https://huggingface.co/google/gemma-4-31B-it [accessed 2026-10-21]
26. Qwen/Qwen3.6-35B-A3B. Hugging Face. URL: https://huggingface.co/Qwen/Qwen3.6-35B-A3B [accessed 2026-10-21]
27. Qwen/Qwen3.8-27B. Hugging Face. URL: https://huggingface.co/Qwen/Qwen3.8-27B [accessed 2026-10-21]
28. Song J, Xu Z, Lu Z, Xiao M, Bi C, Zhang Y, et al. OntologyAligner: ontology-aligned retrieval and hierarchy-guided large language model reranking for biomedical ontology normalization. arXiv. Preprint posted online on September 9, 2026. [doi: 10.48550/arXiv.2609.10055]
29. Vasilevsky NA, Toro S, Matentzoglu N, Flack JE, Mullen KR, Hegde H, et al. Mondo: integrating disease terminology across communities. Genetics. Apr 04, 2026;232(4):iyaf215. [FREE Full text] [doi: 10.1093/genetics/iyaf215] [Medline: 41052288]
30. Ethical review measures for life science and medical research involving human subjects (National Health and Family Planning Commission document no. 4 (2023) [Article in Chinese]. Sichuan University Research Institute. 2024. URL: https://kyy.scu.edu.cn/info/1416/5946.htm [accessed 2026-10-21]
31. Bentegeac R, Le Guellec B, Kuchcinski G, Amouyel P, Hamroun A. Token probabilities to mitigate large language models overconfidence in answering medical questions: quantitative study. J Med Internet Res. Aug 29, 2025;27:e64348. [FREE Full text] [doi: 10.2196/64348] [Medline: 40882190]
32. Song D, Lai S, Wang M, Chen S, Sun L, Wang B. Both text and images leaked! A systematic analysis of data contamination in multimodal LLM. arXiv. Preprint posted online on November 6, 2024. [doi: 10.48550/arXiv.2411.03823]
33. Sheppert AP, Adams B, Sheppert AD, Riley S. Reasoning or reciting? A temporal contamination audit of large language models in clinical medicine. J Am Med Inform Assoc. Sep 01, 2026;33(9):1617-1626. [doi: 10.1093/jamia/ocag069] [Medline: 42281378]
34. Hsieh TC, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, et al. PEDIA: prioritization of exome data by image analysis. Genet Med. Dec 2019;21(12):2807-2814. [FREE Full text] [doi: 10.1038/s41436-019-0566-2] [Medline: 31164752]
35. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, et al. The human phenotype ontology in 2021. Nucleic Acids Res. Jan 08, 2021;49(D1):D1207-D1217. [FREE Full text] [doi: 10.1093/nar/gkaa1043] [Medline: 33264411]
36. Song J, He M, Ren S, Shen B. An explainable dataset linking facial phenotypes and genes to rare genetic diseases. Sci Data. Apr 15, 2025;12(1):634. [FREE Full text] [doi: 10.1038/s41597-025-04922-z] [Medline: 40234471]
37. He M, Song J, Ren S, Zhang Y, Du J, Feng J, et al. FPGDKG 1.0: an integrated facial phenotype-gene-disease knowledge graph for rare disease diagnosis and explanation. IEEE J Biomed Health Inform (Forthcoming). Feb 03, 2026. [doi: 10.1109/JBHI.2026.3659898] [Medline: 41632668]
38. Song J, Xu Z, He M, Feng J, Shen B. Graph retrieval augmented large language models for facial phenotype associated rare genetic disease. NPJ Digit Med. Aug 24, 2025;8(1):543. [FREE Full text] [doi: 10.1038/s41746-025-01955-x] [Medline: 40849403]
39. Abootorabi MM, Zobeiri A, Dehghani M, Mohammadkhani M, Mohammadi B, Ghahroodi O, et al. Ask in any modality: a comprehensive survey on multimodal retrieval-augmented generation. In: Findings of the Association for Computational Linguistics: ACL 2025. Stroudsburg, PA. Association for Computational Linguistics; Jul 2025:16776-16809.
40. Yu Q, Xiao Z, Li B, Wang Z, Chen C, Zhang W. MRAMG-bench: a comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation. In: SIGIR '25: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY. Association for Computing Machinery; 2025:3616-3626.
41. GestaltMatcher Database. URL: https://db.gestaltmatcher.org/ [accessed 2026-10-21]
42. zhelishisongjie/MLLM-Rare-Disease-Diagnosis. GitHub. URL: https://github.com/zhelishisongjie/MLLM-Rare-Disease-Diagnosis [accessed 2026-10-21]

Abbreviations

GMDB: GestaltMatcher Database
LLM: large language model
MLLM: multimodal large language model
MONDO: Mondo Disease Ontology

XSL•FO
RenderX

Edited by I Steenstra; submitted 30.Jun.2026; peer-reviewed by H Yi, K-H Lin, H Li; comments to author 26.Aug.2026; revised version received 15.Oct.2026; accepted 15.Oct.2026; published 02.Nov.2026

Please cite as:
Song J, Xu Z, Xiao M, Bi C, Zhang Y, Zheng X, Li X, Lu Z, Cao Q, Shen B
Diagnosing Rare Genetic Diseases From Facial Images Using Visual Prompting With Multimodal Large Language Models: Diagnostic Accuracy Study
JMIR Med Inform 2026;14:e105860
URL: https://medinform.jmir.org/2026/1/e105860
doi: 10.2196/105860
PMID:

©Jie Song, Zhichuan Xu, Meng Xiao, Cheng Bi, Yuxin Zhang, Xin Zheng, Xiaoran Li, Ziyu Lu, Qiongfang Cao, Bairong Shen. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 02.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX